International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert.v10i12.910
Volume 10, Issue 12, December 2023

Survey Paper

Exploring a Spectrum of Deep Learning Models for Automated Image Captioning: A Comprehensive Survey

Sushma Jaiswal[1*], Harikumar Pallthadka[2], Rajesh P. Chinchewadi[3], Tarun Jaiswal[4]

[1] *Guru Ghasidas Central University, Bilaspur (C.G.) and Post-Doctoral Research Fellow, Manipur International University, Imphal, Manipur, jaiswal1302@gmail.com , https://orcid.org/0000-0002-6253-7327*
[2] *Manipur International University, Imphal, Manipur, vc@miu.edu.in, https://orcid.org/0000-0002-0705-9035.*
[3] *Manipur International University, Imphal, Manipur, Rajesh.cto@miu.edu.in .*
[4] *National Institute of Technology, Raipur, tjaiswal_1207@yahoo.com, https://orcid.org/0000-0003-3963-4548*

**Sushma Jaiswal: jaiswal1302@gmail.com*

Received: 21/10/2023, Revised: 28/11/2023, Accepted: 19/12/2023 Published: 30/12/2023

Abstract: - *Automatic caption generation from images has emerged as a fundamental and challenging problem at the intersection of computer vision and natural language processing. This paper presents a comprehensive survey of the techniques, methodologies, and advancements in the field of automatic caption generation from images. The primary objective is to provide an extensive review of the state-of-the-art models, evaluation metrics, datasets, and applications associated with this domain. The survey begins by elucidating the underlying principles of image feature extraction and caption generation. Various neural network architectures, including Convolutional Neural Networks (CNNs) and recurrent models such as Long Short-Term Memory (LSTM) networks, are discussed in detail. Additionally, the paper explores the integration of attention mechanisms and reinforcement learning strategies to enhance the quality and relevance of generated captions. A thorough examination of evaluation metrics, encompassing both automated and human-centric approaches, is presented to evaluate the generated captions quantitatively and qualitatively. The survey also highlights prominent datasets that have significantly contributed to the advancement of research in this field, facilitating a deeper understanding of challenges and trends. Furthermore, the paper discusses practical applications and real-world use cases where automatic caption generation plays a pivotal role, including accessibility, multimedia indexing, and assistive technologies. The discussion concludes by outlining open challenges and future directions, aiming to inspire further research and innovation in automatic caption generation from images. The aim of this paper is to examine and contrast diverse end-to-end learning frameworks for image captioning, employing established evaluation metrics to comprehend their applicability across different research domains. In addition to the comparative analysis, the paper addresses future challenges in this domain.*

Keywords- *Attention model, CNN, LSTM, Comparative Analysis, image caption.*

---

1. Introduction

Image captioning is a fascinating field at the intersection of computer vision and natural language processing. It involves the development of algorithms and models capable of automatically generating descriptive and coherent textual descriptions for images. The primary objective is to bridge the gap between visual content and human-like understanding by providing contextually meaningful captions that encapsulate the essence of the image. This technology holds immense potential for various applications, including accessibility tools for the visually impaired, enhancing content understanding in search engines, and facilitating a more engaging user experience in social media and e-commerce platforms. The advancements in deep learning, particularly convolutional and recurrent neural networks, have played a pivotal role in propelling the accuracy and effectiveness of image captioning systems.

Finding the qualities of an image is essential to understanding it. The methods employed for this can be

roughly categorized into two groups: (1) Deep machine learning based methods and (2) Traditional machine learning based options.

2. Deep Learning Based Image Captioning Methods

Deep learning-based image captioning involves employing neural networks, often convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for language modeling. CNNs extract meaningful features from images, while RNNs generate descriptive captions based on these features. The "Show and Tell" model [1], which laid the foundation for many subsequent image captioning models by integrating CNNs and RNNs to generate coherent descriptions for images. It's an excellent starting point to delve deeper into the field of deep learning-based image captioning [1].

2.1 Attention Mechanism

The attention mechanism in image captioning is inspired by human visual attention, enabling models to dynamically weigh different regions of the image during caption generation. This approach enhances the relevance of generated words to specific parts of the image, resulting in more accurate and contextually relevant captions. Xu et al. [2] proposed an attention-based image captioning model, revolutionizing the field. The authors introduced an attention mechanism that allows the model to focus on different parts of the image while generating words in the caption. It serves as a cornerstone for understanding and implementing attention mechanisms in image captioning [2].

2.1.1. Categories of Attention Mechanisms

The attention Mechanism has a complex cognitive ability, stemming from the human brain. Attention is the ability of self-selection. In the neural network model, the attention mechanism allows the neural network to focus on its sub inputs to find the specific feature of an image. We may categories this attention mechanism in following categories:

2.1.1.1 Soft Attention

Soft attention mechanisms dynamically allocate weights to various regions of the input image, allowing the model to focus on the most relevant parts during caption generation. This leads to more descriptive and contextually accurate captions by aligning the generated words with specific image features. Bahdanau et al. [48] introduced the attention mechanism in the context of neural machine translation. Although not directly focused on image captioning, it laid the foundation for the application of attention mechanisms in various domains, including image captioning. This work demonstrates the early conceptualization of the soft attention mechanism, which has been influential in subsequent image captioning research.

2.1.1.2 Hard Attention

Hard attention mechanisms in image captioning involve making discrete decisions to attend to specific regions of the image during caption generation. These mechanisms produce more interpretable results as they select only one region at a time, improving the model's ability to generate meaningful and relevant captions. A hard attention mechanism that uses a 'visual sentinel' to decide whether to focus on the image or generate a new word in the caption. This approach enables the model to make discrete attention decisions, leading to improved interpretability in image captioning. It serves as a foundational work in the realm of hard attention mechanisms for image captioning [3].

2.1.1.3 Multi Head Attention

Multi-head attention in image captioning involves employing multiple attention mechanisms, each focusing on different parts of the image. This approach enables the model to capture diverse and complex features, resulting in more comprehensive and contextually rich captions. The authors [4] proposed a multi-head attention mechanism for image captioning. By combining bottom-up and top-down attention, this approach attends to multiple regions of the image, allowing the model to generate more informative and contextually relevant captions. It serves as a fundamental reference for understanding the integration of multi-head attention in image captioning.

2.1.1.4 Scaled Dot-Product Attention

Scaled Dot-Product Attention is a technique used to calculate attention scores between two sets of vectors. It involves scaling the dot product of the query and key vectors by the square root of the dimension of the key vectors. This operation provides a more stable and well-scaled computation of attention scores, which is crucial in attention-based models. The paper "Attention is All You Need"[5] introduced the Transformer model, which extensively uses the Scaled Dot-Product Attention mechanism. While the focus is on natural language processing and not image captioning, understanding this foundational mechanism is crucial for grasping the core concepts of attention-based models, which are prevalent in both domains.

2.1.1.5 Global Attention

Global attention mechanisms in image captioning involve considering the entire image during the caption generation process. Unlike traditional attention mechanisms, global attention ensures that every part of the image contributes to the caption, promoting a more holistic understanding and description of the visual content. The

global attention [6] concept is fundamental for image captioning as well, providing insights into utilizing global information from the image to generate coherent and comprehensive captions. Understanding this mechanism is crucial for leveraging global attention in image captioning models.

2.1.1.6 Local Attention

Local attention mechanisms in image captioning enable the model to attend selectively to particular regions of the image. This approach allows the model to align the generated words with specific visual features, enhancing the relevance and informativeness of the generated captions. Luong et al. [7] work laid the foundation for applying local attention in various domains, including image captioning. Understanding the local attention mechanism is essential for leveraging region-specific information in image captioning models [7].

2.1.1.7 Adaptive Attention

Adaptive attention mechanisms in image captioning facilitate dynamic adjustments of attention weights, allowing the model to effectively focus on specific regions of the image. This adaptability enhances the alignment of generated captions with relevant visual features, resulting in more meaningful and coherent descriptions. This approach [8] dynamically adapts the attention weights based on the generated words, allowing the model to refine its focus on relevant image regions during caption generation. Understanding adaptive attention mechanisms is crucial for improving the flexibility and effectiveness of image captioning models.

2.1.1.8 Semantic Attention

Semantic attention in image captioning involves directing the model's attention to semantically important regions of the image. This helps to generate captions that are more aligned with the high-level concepts present in the image, enhancing the overall relevance and informativeness of the generated captions. This approach [9] focuses on semantically relevant regions of the image, allowing the model to generate captions that are conceptually closer to the content of the image. Understanding semantic attention is crucial for improving the quality and relevance of image captions in various applications.

2.1.1.9 Areas of Attention

In image captioning, "areas of attention" refer to specific regions or parts of the image that the model focuses on during the caption generation process. This approach aims to provide more relevant and detailed captions by aligning the generated text with the salient features and regions of the image. The authors [10] introduced a mechanism where the model attends to different areas of the image while generating captions. This approach significantly improves the relevance and informativeness of the generated captions. Understanding and utilizing areas of attention is crucial for enhancing the performance of image captioning models.

2.1.1.10 Deliberate Attention

Deliberate Attention creates image captions in accordance with people's habits.

2.2 Transformer-based Models

In "Image Transformer," Dosovitskiy et al. [11] proposed a transformer-based model that directly processes images as sequences of patches, enabling parallel processing and capturing long-range dependencies within the image. This approach achieved competitive performance on image generation tasks compared to convolutional neural network (CNN)-based models. The transformer architecture, known for its self-attention mechanisms and ability to capture contextual relationships across input elements, was harnessed to model image data. By dividing images into patches and treating them as a sequence, the model could learn to generate meaningful representations and effectively generate captions for images.

2.3 Traditional Approach in Image Captioning

This paper by Farhadi et al. [12] presents an early approach to image captioning. The authors propose a method to generate sentences describing images by leveraging object and attribute detectors, aiming to automatically generate captions that de-scribe the content of the image. The paper by Kulkarni et al. [13] focuses on generating simple image descriptions using statistical language models. It aims to describe images in a basic and straightforward manner, laying the foundation for later advancements in image captioning. The paper by Elliott, Rottenberg, and Stankiewicz [14] presents an early AI system that utilizes a set of rules to describe visual information, demonstrating a pioneering effort in generating captions based on a structured understanding of images.

2.4 Template-based Image Captioning

The paper by Ordonez, Kulkarni, and Berg [15] introduces the "Im2Text" approach, which generates image descriptions using predefined templates. The tem-plates consist of sentence structures, and placeholders are filled based on detected objects and their attributes in the image. This work represents an early exploration of template-based image captioning, providing a foundation for subsequent research in this domain. Please note that advancements in deep learning have largely shifted the field towards end-to-end trainable models, but template-

based approaches remain important for certain applications and as a reference point for newer techniques.

2.5 Retrieval-Based Image Captioning

The paper by Hodosh, Young, and Hockenmaier [16] presents a retrieval-based approach to image description generation, framing the task as a ranking problem. The model ranks a set of pre-existing captions based on their relevance to the input image, and the highest-ranked caption is selected as the description for the input image. This work laid the foundation for retrieval-based image captioning and provided insights into evaluation metrics and methodologies for this approach. It's important to note that while retrieval-based methods have their merits, recent advancements in deep learning have led to a significant shift towards end-to-end trainable models for image captioning.

2.6 Novel Image Caption Generation

The paper by Lu et al. [17] introduced "Neural Baby Talk," a novel approach that generates detailed and fine-grained captions for images. It leverages object detection results and uses a neural network to learn the relationships between objects in an image and generate a caption that describes the scene in a coherent and informative manner. This work represents a step towards generating more nuanced and detailed image captions by considering relationships between objects and leveraging deep learning techniques. The field of novel image caption generation continues to evolve, with advancements in models, architectures, and evaluation methods. It's important to explore the latest research to stay updated with the most recent developments.

3. Datasets

Image captioning databases usually comprise of image pairings with matching captions. These datasets are frequently utilized in the development and assessment of machine learning models that produce textual descriptions for photographs. The MS COCO (Microsoft Common Objects in Context) [18], Flickr30k [19], Flickr8k [19], Visual Genome [20], AI2 Thor [21], VGG Image Annotator [22] and COCO-Stuff [23] dataset is a popular resource for image captioning. brief explanation given in Table 1.

4. Evaluation Metrics

An overview of various popular evaluation metrics for machine learning and natural language processing tasks is provided below (See Table 2). Every statistic has a distinct function and sheds light on various facets of the model's performance

5. Comparison on benchmark datasets and common evaluation metrics

Comparison of some benchmark datasets commonly used in image captioning on the following basis (See Table 3): It is crucial to select a dataset according to the particular specifications and features of the visual captioning assignment at hand. Every dataset has distinct qualities of its own, and the selection is based on various aspects like the required level of complexity, variety of scenes, and annotation detail. To guarantee the stability and applicability of their Image captioning algorithms, researchers frequently employ a variety of datasets.

The model name suggests that multilayer CNN and multilayer LSTM were utilized. This method allows for the development of multilevel semantics for language and vision using CNN and LSTM, respectively, which can then be fused together to form the final caption. As a result, the model improves network capacity and yields better results [30]. While LSTM is used to generate sentences, CNN is used to generate features. For improved semantic information, bi-directional LSTM refers to two distinct LSTMs, one for the historical context and another for the future context. These models have yielded more semantically accurate captions. The evaluation metrics using this model on several datasets are displayed in the table. To increase the model's efficiency, visual attention might be incorporated in addition to Bi-LSTM [31]. A set of image regions with prominent features is provided by the bottom-up mechanism. This model uses a Recurrent-Convolution Neural Network (R-CNN) to implement the Bottom Up portion. The model is made efficient by using a top-down technique to forecast the attention distribution across image regions. The final feature vector is generated by taking the weighted average of all the features in the image. In comparison to earlier suggested models, the model is yielding greater Blue scores and CIDEr [4].

Mao et al. [45] obtained 0.565, 0.386, 0.256, and 0.170 on BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively, as indicated in *Table 4* on Flickr8k. The scores for the Flickr30k dataset are higher than the Flickr8k dataset, at 0.600, 0.410, 0.280, and 0.190, respectively. The MSCOCO dataset yielded the highest scores. Larger datasets yield superior results because they contain more data, provide a more thorough depiction of different scenarios, are more complicated, and have their own natural context. And Flickr30k datasets, the technique maps textual and image elements using visual space. Mao et al. map text features to image features using multimodal space. However, Jia and colleagues map using visual space. Additionally, the technique makes use of an encoder-decoder design, allowing it to dynamically direct the decoder portion. As a result, this approach outperforms Mao et al.

Additionally, Xu et al. [10] outperform on the MSCOCO dataset. This approach fared better than Jia et al. [46] and Mao et al. [45], This is primarily because it employs an attention mechanism that restricts focus to the image's pertinent objects. Semantically rich captions can

be produced via the semantic concept-based methods. A semantic concept-based approach for captioning images was presented by Wu et al. [47]This approach uses the image to first anticipate the qualities of various objects, and then it appends these semantically important attributes to the captions. The method outperforms every other method listed in Table 2 in terms of performance.

The outcomes of attention-based techniques on the MSCOCO dataset are displayed in Table 2. Stochastic hard attention by Xu et al. [2] outperformed deterministic soft attention in terms of results. Yet, Jin et al. [46] superior performance was due to its ability to adjust its focus according to the particular context of each scene.

Only BLEU-4 and METEOR scores—which are higher than the previously described methods—are displayed by Wu et al. [47] and Pedersoli et al. [39]. Wu et al.'s method combines a review procedure with an attention mechanism. Every time a step in the review process is completed, the concentrated attention is checked and updated as needed. Compared to earlier attention-based techniques, this process aids in achieving superior outcomes. Rather than using an LSTM state, Pedersoli et al. [39] suggest an alternative attention mechanism that directly maps the focused image regions with the caption words. Because of this behavior, the approach outperforms the other attention-based methods shown in Table 2.

Both GAN-based and reinforcement learning (RL)-based techniques are gaining traction. "Other Deep Learning-based Image Captioning" is the moniker we give them. Table 2 displays the outcomes of this group's methods. The approaches yield no results on widely used assessment metrics. They might, nevertheless, come up with the image's descriptions on their own. In Shetty et al. [40] image captioning technique, adversarial training was used. This technique can provide a variety of captions. When comparing the methods that use maximum likelihood estimate, the captions are less biased when they use the ground-truth captions. Ren et al. [41] suggested a method that can anticipate all possible following words for the current word in the current time step in order to take advantage of RL's benefits. This system aids in producing captions that are more correct in terms of context. The Discriminator and Generator of a GAN are comparable to the Actor-critic of RL. Nevertheless, neither the reviewer nor the performer knew anything about data at the start of the course. An actor-critic based approach for captioning images was presented by Zhang et al. [43] Compared to other reinforcement learning-based techniques, this method can produce more accurate captions and can forecast the final captions at an early stage.

6. Challenges and Future Directions

The interpretation of ambiguous visual elements, handling long-range dependencies in images, handling uncommon or unseen concepts, integrating multi-modal information, describing fine-grained details, capturing context awareness, managing multiple objects within an image, choosing appropriate evaluation metrics, generating diverse captions, and optimising for real-time processing are just a few of the challenges involved in developing effective image captioning models. To overcome these obstacles, a comprehensive strategy combining developments in natural language processing and computer vision is needed to make sure that models function effectively in real-world settings, generalise successfully, and generate comprehensive, contextually aware descriptions.

6.1 Challenges

When investigating a range of deep learning models for automated image captioning, scientists could run into a number of difficulties. The following difficulties might be looked into in the framework of such a survey:

- **Diversity and Creativity:** Generating diverse and creative captions for images.
- **Fine-Grained Details:** Describing intricate image details and relationships accurately.
- **Multimodal Integration:** Effectively using diverse inputs like text, audio, and video.
- **Context Sensitivity:** Generating context-aware captions for user-specific applications.
- **Ambiguity Handling:** Addressing ambiguity and uncertainty in image interpretation.
- **Data Bias and Fairness:** Ensuring fairness and inclusivity in generated captions.

6.2 Future Directions in Image Captioning

Image captioning is expected to progress in a number of important areas in the future. First, models are expected to advance in image understanding by becoming more proficient in managing intricate scenarios and comprehending minute details. To improve the naturalness and relevancy of the descriptions, there will also need to be a greater emphasis on creating diverse and contextually rich captions. The handling of long-range relationships and the coherence of generated captions may be further improved by integrating attention processes and reinforcement learning approaches. Building more reliable and broadly applicable models will be aided by the investigation of bigger and more varied datasets, including those containing uncommon ideas. Finally, studies may focus more on practical applications, like captioning images in dynamic settings, allowing models to function well in real-time situations, and enhancing their applicability in a range of contexts.

- **Enhancing Diversity:** Developing models that produce a wider range of caption styles.
- **Fine-Grained Recognition:** Advancing image understanding for detailed captions.
- **Multimodal Fusion:** Improving integration of diverse input modalities.
- **Contextual Adaptation:** Creating captions sensitive to context and user preferences.
- **Ambiguity Management:** Incorporating mechanisms to handle ambiguity.

- **Bias Mitigation:** Reducing biases in training data and generated captions.

Through tackling these obstacles, scholars may enhance a more all-encompassing comprehension of the potential and constraints of deep learning models for automatic image captioning.

Table 1. Dataset Description

Dataset	**Description**	**Content**	**Annotations**	**Challenges**	**Usage**
MS COCO [18]	Large-scale dataset for object recognition, segmentation, and captioning.	Over 200,000 images with detailed captions.	Object instance segmentation masks, object keypoints, per-object captions.	Annual challenges for benchmarking. Widely adopted in computer vision research.	Image captioning models, object recognition, segmentation.
Flickr30k [19]	Collection of 30,000 images from Flickr, each with five human-generated captions.	30,000 images	Five human-generated captions per image.	Benchmarking and comparing image captioning models.	Image captioning research.
Flickr8k [19]	Similar to Flickr30k but with 8,000 images.	8,000 images	Five human-generated captions per image.	Benchmarking and comparing image captioning models.	Image captioning research.
Visual Genome [20]	Extensive dataset with over 100,000 images. Contains object-level, stuff-level, and region-level annotations.	Over 100,000 images	Object-level, stuff-level, and region-level annotations.	Versatile for various computer vision tasks.	Object recognition, scene understanding.
AI2 Thor [21]	Focuses on indoor scenes with images from 3D environments. Provides dense captions and annotations for object relationships.	Images from 3D environments	Dense captions, annotations for object relationships.	Specific focus on indoor scenes.	Indoor scene understanding.
VGG Image Annotator [22]	Allows users to annotate images with regions and keypoints. Users can define their annotation tasks.	User-defined annotation tasks	Annotations for regions and keypoints.	Versatile for various annotation tasks.	Custom annotation tasks, versatile applications.
COCO-Stuff [23]	An extension of MS COCO with additional annotations for 91 stuff classes.	Extension of MS COCO	Additional annotations for 91 stuff classes.	Extends MS COCO for more detailed stuff annotations.	Scene understanding, stuff recognition.

Table 2. Evaluation Metrics

Metric	**Description**	**Advantages**	**Considerations**
BLEU (Bilingual Evaluation Understudy) [24]	Measures precision of n-grams in generated caption compared to reference captions.	Simple and easy to calculate.	May not capture quality and fluency well.
METEOR (Metric for Evaluation of Translation with Explicit ORdering) [25]	Considers precision, recall, and alignment between generated and reference words. Incorporates stemming and synonymy.	Accounts for variations and synonyms.	Sensitive to tokenization and preprocessing.
ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [26]	Evaluates overlap of n-grams and word sequences between generated and reference captions.	Emphasizes recall.	Limited in assessing fluency and coherence.
CIDEr (Consensus-based Image Description Evaluation) [27]	Measures consensus between multiple reference captions and the generated caption, considering similarity at the word and phrase level.	Addresses some BLEU and METEOR limitations.	Sensitive to the number of reference captions.
SPICE (SPecific Image Caption Evaluation) [28]	Evaluates semantic content by comparing generated captions to reference captions in terms of object, attribute, and relationship	Focuses on semantic content.	Requires additional processing of references.

	triplets.		
MUTT (Metric for Unsupervised Image-to-Text Transfer) [29]	Designed to evaluate caption quality without relying on ground truth references, considering informativeness, fluency, and diversity of generated captions.	Addresses challenge of evaluating without references.	Relatively new, needs further validation.

Table 3. Comparison of some benchmark datasets commonly used in image captioning

Aspect	**Benchmark Datasets**
Purpose	Datasets designed specifically for image captioning tasks.
Domains	Diverse domains, covering everyday scenes, objects, and activities. Examples include indoor scenes, outdoor scenes, and images from various sources like Flickr.
Size	Varies from moderate-scale datasets with thousands of images to large-scale datasets with hundreds of thousands of images.
Annotations	Annotations typically include multiple human-generated captions per image, providing diverse descriptions for each visual input. Annotations may also include object segmentation masks, keypoints, and other detailed information.
Common Datasets	- MS COCO (Microsoft Common Objects in Context): Large-scale dataset with diverse scenes and detailed annotations. Flickr30k: Collection of 30,000 images from Flickr with human-generated captions. Flickr8k: Similar to Flickr30k but with 8,000 images. Visual Genome: Extensive dataset with object-level, stuff-level, and region-level annotations. AI2 Thor: Focuses on indoor scenes with 3D environments. COCO-Stuff: Extension of MS COCO with additional annotations for stuff classes.
Challenges	MS COCO hosts annual challenges for benchmarking image captioning algorithms. Challenges often focus on evaluating the diversity, accuracy, and fluency of generated captions.
Usage	Widely used for training, validating, and benchmarking image captioning models. Commonly adopted in the computer vision community for assessing the quality of generated captions.

Table 4. Comparative analysis of various models for image captioning

Methods	Dataset	Evaluation Metrics						
		B-1 (%)	**B-2 (%)**	**B-3 (%)**	**B-4 (%)**	**METEO R (%)**	**R-L (%)**	**CIDEr**
Vinyals et al. 2015 [1]	Flickr8K	0.63	-	-	-	-	-	-
	Flickr30K	0.66	-	-	-	-	-	-
	MSCOCO	-	-	-	0.277	0.237		0.855
Xu et al. 2015 [2]	Flickr8K	0.67	0.448	0.299	0.195	0.203		
	Flickr30K	0.669	0.439	0.296	0.199	0.1846		
	MSCOCO	0.718	0.504	0.357	0.25	0.2304		
Xiao et al. 2018 [30]	Flickr8K	0.651	0.47	0.326	0.205	0.205	-	-
	Flickr30K	0.654	0.468	0.329	0.231	0.193	-	-
	MSCOCO	0.731	0.563	0.426	0.323	0.258	0.538	1.001
WANG et at.2018 [31]	Flickr8K	0.669	0.484	0.333	0.228	-	-	
	Flickr30K	0.636	0.448	0.304	0.205	-	-	
	MSCOCO	0.687	0.509	0.364	0.258	0.229		0.739
Anderson et al.2019 [4]	MSCOCO	0.772	-	-	0.362	0.270	0.564	1.135
Maofu et al. 2019 [32]	AICICC	0.741	0.614	0.509	0.423	0.35	0.602	1.236
Yuting et al. 2019 [33]	MSCOCO	0.724	0.558	0.421	0.318	0.254	0.536	1.012
YANG et al. 2020 [9]	MSCOCO	0.781	-	-	0.367	0.285	0.584	1.192

Xiaodan et al. 2019 [34]	Flickr8K	0.598	0.408	0.275	0.184	-	-	-
	Flickr30K	0.592	0.391	0.257	0.17	-	-	-
	MSCOCO	0.675	0.498	0.364	0.269	0.226	0.80	0.499
Junhao et al. 2020 [35]	MSCOCO	0.813	0.654	0.507	0.385	0.285	0.588	1.235
Eric et al. 2019 [36]	Flickr8K	-	-	0.477	0.334	0.231	0.469	167.5
	Flickr30K	-	-	0.52	0.391	0.26	0.51	2.091
	MSCOCO	-	-	0.51	0.34	0.248	0.565	1.183
	MSCOCO (Inception V3)	0.695	0.518	0.38	0.277	0.235	-	0.894
Ruifan et al. 2020 [37]	Stanford Image paragraph dataset	0.416	0.244	0.143	0.86	0.156	-	0.174
Xu et al. 2015 [2], soft Xu et al. 2015 [2], hard	MSCOCO	0.707 0.718	0.492/ 0.504	0.344/ 0.357	0.243/ 0.250	0.239/ 0.230	-	-
Wu et al. 2016 [38]	MSCOCO	-	-	-	0.290	0.237	-	0.886
Pedersoli et al. 2017 [39]	MSCOCO	-	-	-	0.307	0.245	-	0.938
Shetty et al. 2017GAN [40]	MSCOCO	-	-	-	-	0.239	-	-
Ren et al. 2017RL [41]	MSCOCO	0.713	0.539	0.403	0.304	0.251	0.525	0.937
Zhang et al. 2017RL [42]	MSCOCO	-	-	-	0.344	0.267	0.558	1.162
Lu et al. 2017 [3]	MSCOCO	0.742	0.580	0.439	0.332	0.266	-	1.085
Gan et al. 2017 [42]	MSCOCO	0.741	0.578	0.444	0.341	0.261	-	1.041
Zhang et al. 2017 [43]	MSCOCO	-	-	-	0.344	0.267	0.558	1.162
Rennie et al. 2017 [44]	MSCOCO	-	-	-	0.319	0.255	0.543	1.06
Wu et al. 2018 [47]	Flickr8k	0.740	0.540	0.380	0.270	-	-	-
	Flickr30k	0.730	0.550	0.400	0.280	-	-	-
	MSCOCO	0.740	0.560	0.420	0.310	0.260	-	-
Jia et al. 2015 [46]	Flickr8k	0.647	0.459	0.318	0.216	0.201	-	-
	Flickr30k	0.646	0.466	0.305	0.206	0.179	-	-
	MSCOCO	0.670	0.491	0.358	0.264	0.227	-	-
Xu et al. 2015 [2]	Flickr8k	0.670	0.457	0.314	0.213	0.203	-	-
	Flickr30k	0.669	0.439	0.296	0.199	0.184	-	-
	MSCOCO	0.718	0.504	0.357	0.250	0.230	-	-
Mao et al. 2015 [45]	Flickr8k	-	0.565	0.386	0.256	0.170	-	-
	Flickr30k	0.600	0.410	0.280	0.190	-	-	-
	MSCOCO	0.670	0.490	0.350	0.250	-	-	-

7. **Conclusion**

In this paper we provide a comprehensive overview and analysis of the diverse range of deep learning models applied to automated image captioning. The survey delves into the advancements and methodologies within this domain, aiming to shed light on the progression of Techniques and their implications. The survey systematically categorizes and discusses various deep learning models, ranging from traditional architectures to state-of-the-art approaches, each contributing uniquely to image captioning. By exploring the evolution of these models, the survey highlights the significant shift from conventional handcrafted feature-based methods to data-driven, end-to-end trainable models. The critical insights offered in this survey emphasize the importance of attention mechanisms, multimodal fusion, recurrent neural

networks, and transformer-based architectures in significantly enhancing the quality and relevance of generated captions.

Additionally, the exploration of evaluation metrics and benchmark datasets underscores the necessity of robust evaluation strategies to measure the effectiveness of image captioning models accurately. Furthermore, the survey addresses the ongoing challenges and potential future directions in image captioning, such as generating diverse and coherent captions, handling fine-grained details, and incorporating context-awareness. These challenges present exciting opportunities for future research and innovation in the field. Overall, "Exploring a Spectrum of Deep Learning Models for Automated Image Captioning: A Comprehensive Survey" serves as a valuable resource for researchers, practitioners, and enthusiasts in the computer vision and natural language processing communities, offering a comprehensive understanding of the landscape of deep learning models for image captioning and paving the way for future advancements in this captivating domain.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages. 3156-3164, 2015.

[2] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In International Conference on Machine Learning (ICML), pages 2048-2057, 2015.

[3] Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3242-3250, 2017.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6077-6086, 2018.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is All You Need. In Advances in Neural Information Processing Systems, pages 30-38, 2017).

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh (2017). Hierarchical Question-Image Co-Attention for Visual Question Answering. In Advances in Neural Information Processing System, pages 289-298, 2017.

[7] Minh-Thang Luong, Hieu Pham, Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),pages 1412-1421, 2015.

[8] Huang, Lun, Wenmin Wang, Yaxian Xia and Jie Chen. "Adaptively Aligned Image Captioning via Adaptive Attention Time." *ArXiv* abs/1909.09060, 2019: n. pag.

[9] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo. Image Captioning with Semantic Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4651-4659, 2016.

[10] Yang Li, Lukasz Kaiser, Samy Bengio, Si Si. Area Attention. Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), pages 20037-20047, 2020.

[12] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier & David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part V (ECCV), pages 15-29, 2010.

[13] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In Proceedings of the 24th CVPR, pages 1609-1616, 2011.

[14] Elliott, R., Rottenberg, A., & Stankiewicz, B. An AI system that describes its understanding of visual information. In Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI), pages 1483-1489, 1993.

[15] Vicente Ordonez, Girish Kulkarni, Tamara Lee Berg. Im2Text: Describing images using 1 million captioned photographs. In Advances in Neural Information Processing Systems (NIPS), pages 1143-1151, 2011.

[16] Micah Hodosh, Peter Young, Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47, 853-899, 2013.

[17] Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh. Neural Baby Talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7219-7228, 2018

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár. Microsoft COCO: Common Objects in Context. arXiv preprint arXiv:1405.0312, 2014.

[19] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic

inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78, 2014.

[20]Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. arXiv preprint arXiv:1602.07332, 2016.

[21]Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv preprint arXiv:1712.05474, 2017.

[22]Abhishek Dutta, Gupta, A., Andrew Zisserman. VGG Image Annotator (VIA): A Simple and Efficient Tool for Annotation of Images and Videos. Proceedings of the European Conference on Computer Vision (ECCV), 242-257, 2016.

[23]Holger Caesar, Jasper Uijlings, Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. arXiv preprint arXiv:1612.03716, 2018.

[24]Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting on association for computational linguistics, 311-318, 2002.

[25]Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 65-72, 2005.

[26]Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL-04 workshop, 74-81, 2004.

[27]Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh. CIDEr: Consensus-based image description evaluation. Proceedings of the IEEE conference on computer vision and pattern recognition, 4566-4575, 2015.

[28]Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. European conference on computer vision, 382-398, 2016.

[29]Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, Yejin Choi.Simulating Action Dynamics with Neural Process Networks. arXiv preprint arXiv:1805.09921, 2018.

[30]Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan. Deep Hierarchical Encoder-Decoder Network for Image Captioning. IEEE Transaction on Multimedia, Apr-2018.

[31]Cheng Wang, Haojin Yang, and Christoph Meinel. Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning. ACM Trans. Multimedia Comput. Commun. Appl., Vol. 14, No. 2s, Article 40. April-2018.

[32]Maofu Liua,b, Lingjun Li, Huijun Hu, Weili Guan, Jing Tian. Image caption generation with dual attention mechanism. Information Process and Management - 57, Elsevier, Nov-2019.

[33]Yuting Su, Yuqian Li, Ning Xu, An-An Liu. Hierarchical Deep Neural Network for Image Captioning. Neural processing letters, Springer Nature, 2019.

[34]Xiaodan Zhang, Shengfeng He, Xinhang Song, Rynson W.H. Lau, Jianbin Jiao, Qixiang Ye. Image captioning via semantic element embedding. Neurocomputing, Elsevier, June-2019.

[35]Junhao Liu, Kai Wang, Chunpu Xu, Zhou Zhao, Ruifeng Xu, Ying Shen, Min Yang. Interactive Dual Generative Adversarial Networks for Image Captioning. Association for the Advancement of Artificial Intelligence, 2020.

[36]Eric ke wang, Xun zhang, Fan wang, Tsu-yang wu, and Chien-ming chen, Multilayer Dense Attention Model for Image Caption, IEEE Access, June-2019.

[37]Ruifan Li, Haoyu Liang, Yihui Shi, Fangxiang Feng, Xiaojie Wang.Dual-CNN: A Convolutional language decoder for paragraph image captioning. Neurocomputing, Elsevier, Feb-2020.

[38]Yang, Zhilin & Yuan, Ye & Wu, Yuexin & Salakhutdinov, Ruslan & Cohen, William. Encode, Review, and Decode: Reviewer Module for Caption Generation, 2016.

[39]Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. "Areas of Attention for Image Captioning". In: Proceedings of the IEEE International Conference on Computer Vision, pages 1251–1259, 2017.

[40]Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In: IEEE International Conference on Computer Vision (ICCV), pages 4155–4164, 2017.

[41]Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep Reinforcement Learningbased Image Captioning with Embedding Reward. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1151–1159, 2017.

[42]Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1141–1150, 2017.

[43]Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. In: 31st Conference on Neural Information Processing Systems. 2017.

[44]Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Selfcritical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1179–1195, 2017.

[45]Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with

multimodal recurrent neural networks (m-rnn). In: International Conference on Learning Representations (ICLR). 2015.

[46] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision, pages 2407–2415, 2015.

[47] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. "Image captioning and visual question answering based on attributes and external knowledge". In: vol. 40. 6. IEEE, pages 1367–1381, 2018.

[48] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio.Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473, 2014.